Inter-variable correlation prediction with fuzzy connected-triples

Li, Zhenpeng; Shang, Changjing; Shen, Qiang

Published in:
Soft Computing

DOI:
10.1007/s00500-018-3427-z

Publication date:
2018

Citation for published version (APA):
Li, Z., Shang, C., & Shen, Q. (2018). Inter-variable correlation prediction with fuzzy connected-triples. *Soft Computing*, *22*(21), 7059-7072. https://doi.org/10.1007/s00500-018-3427-z

Soft Computing
https://doi.org/10.1007/s00500-018-3427-z

FOCUS

Inter-variable correlation prediction with fuzzy connected-triples

Zhenpeng Li[1] · Changjing Shang[1] · Qiang Shen[1]

Abstract
Identification of hidden relationships between domain attributes from different data sources is of great practical significance and forms an emerging field in data mining. However, currently there seldom exist any systematic methods that can effectively handle this problem, especially when dealing with imprecisely described associations. In this paper, a novel data-driven approach for inter-variable correlation prediction is proposed by exploiting the concept of connected-triples. The work is implemented with the use of fuzzy logic. Through the exploitation of link strength measurements and fuzzy inference, the job of detecting similar or related variables can be accomplished via examining link relation patterns within and across different data sources. Empirical evaluation results are discussed, revealing the potential of the proposed work in predicting interesting attribute relations, while involving simple computation mechanisms.

Keywords Connected-triple · Fuzzy inference · Inter-variable correlation · Link prediction

1 Introduction

Over the past decade, data has been playing an increasingly significant role in daily life. Fortunately, the rapid growth of computing capability and computational techniques has to a certain extent enabled the handling of such large amount of data. A wide range of potentially effective approaches exist, including the method of social network analysis (SNA) that has been increasingly gaining popularity in solving real-world problems.

Social networks are fundamentally social structures including actors and relationships amongst them (Wasserman and Faust 1994). These networks can be conveniently represented by employing vertices and links. The links show types of relationship amongst the vertices including kinship, friendship, collaborations and any other interactions between the people, namely the vertices in such a network (Newman and Park 2003). In particular, it is widely applied in recommendation systems for information retrieval, helping search for new friends (Aiello et al. 2012) and potential business collaborators (Akcora et al. 2011; Mori et al. 2012; Wu et al. 2013), finding domain experts or co-authors in academic fields (Pavlov and Ichise 2007). Obviously, the concept of SNA models can be generalised. They are not only restricted to the use in networks concerning human beings, but also can be utilised to depict and analyse the structures in a wide variety of problem domains.

Recently, SNA has become an important and effective technique in the study of biology, economy and other cross-disciplines (Wasserman and Galaskiewicz 1994). For instance, in bioinformatics, SNA is adjusted to present gene expression networks (Almansoori et al. 2012), describing protein–protein interactions (Franceschini et al. 2012). In politics research, SNA is adopted to conceptualise a policy-making process as a network of political actors (Varone et al. 2017). In public health care, SNA is employed to assess factors contributing to the service, the care process and the patient outcome (Bae et al. 2015). In project management, SNA is utilised to measure the correlations amongst stakeholders, process-related values and outcome-related values (Zheng et al. 2016) In E-commerce, SNA is equipped for providing interesting items in online shopping (Akcora et al. 2011). Last but not the least, in the field of national defence

Communicated by F. Chao, Q. Zhang.

✉ Changjing Shang
cns@aber.ac.uk

Zhenpeng Li
zhl6@aber.ac.uk

Qiang Shen
qqs@aber.ac.uk

[1] Department of Computer Science, Institute of Mathematics, Physics and Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK

Published online: 01 August 2018

and public security, SNA is assembled for terrorism and insurgency detection (Boccaletti et al. 2006; Shen and Boongoen 2012a, b; Zech and Gabbay 2016), money laundering prevention (Colladon and Remondi 2017) and abnormal telecommunication surveillance (Huang and Lin 2009).

In SNA, link prediction is one of the most salient and challenging tasks. It is particularly difficult to perform the discovery of missing or developing links in a certain network of interest (Liben-Nowell and Kleinberg 2007). However, link prediction is very useful to help: infer the underlying complete network (from partially observed structures) (Marchette and Priebe 2008; Kim and Leskovec 2011), understand the evolution of networks (Bringmann et al. 2010; Barabâsi et al. 2002) and predict hyperlinks in heterogeneous social networks (Zhu et al. 2002). Traditionally, most of the approaches for detecting unobserved links are based on topological information, including neighbour-based metrics, path-based metrics and random walk-based metrics (Wang et al. 2015). Recent studies have extended such classical metrics by adding weights to the existing links within a topological graph in response to the information obtained from explicitly related sources (Lü and Zhou 2010). Also, probabilistic methods have been proposed to handle different forms of link prediction under uncertainty (Martínez et al. 2017). However, typical existing approaches (including all discussed above) are set for a specific problem within a local scope, dealing with the information coming from a single data source.

Addressing the task of link prediction, the use of connected-triples has an intuitive appeal. A connected-triple is a graph representation formed by three vertices and two undirected edges, with each edge connecting two distinct vertices out of the three, respectively. A network constructed with such connected-triples offers a potentially effective mechanism for link prediction, particularly when any given information content is obtained from different data sources where parts of the information overlap. Inspired by this observation, unlike previous research that focussed on identifying links between objects or entities in a specific region, this paper presents an innovative piece of work that is driven by the interests in searching for links between variables extracted from different data/information sources, through the introduction and exploitation of fuzzy connected-triple.

The potential underlying links between variables or entities collected from different sources are usually hidden, not obvious or even difficult to be discovered, making the task of link prediction from such data sources a challenge. Traditionally, this type of work has generally been handled by human experts. Thus, designing and implementing a predicting method which learns from human logical reasoning will be helpful to automate such prediction processes, especially when facing large and diverse data sources. Practically, when describing a link or a set of links, linguistic terms such as "strong", "medium" and "weak" are natural adjectives to depict the link strength rather than crisp numerical values (that are typically utilised in conventional connected-triple models). In addition, common knowledge such as "if A has a strong link to B, and B has a strong link to C, then A may have a strong link to C" perfectly matches human logical thinking. It is to reflect such intuitions, fuzzy logic is adopted in the present work to serve as the basis upon which to develop a multi-source link prediction model. Such link prediction problems are obviously of general interest in many data mining applications.

Overall, this paper presents two major contributions to knowledge: (1) It proposes a novel approach to determining the correlation between attribute variables from distinct datasets with different entity references. (2) It proposes a fuzzy link prediction model which radically departs from conventional crisp representation of connected-triple-based link detection, resulting in models that resemble human inference and facilitate interpretability. The rest of this paper is arranged as follows. Section 2 introduces the proposed architecture for the development of a fuzzy connected-triple system for link prediction, describing details on model construction, link measures and inference procedures. Section 3 exhibits the results of empirical evaluation, supported by comparative studies with alternative predicting methods. Section 4 concludes the paper with outlook for further development.

2 Predicting system

This section presents the proposed general framework for developing a system that predicts link strengths with data from multiple sources. It describes the system's components and their associated time complexity analyses.

2.1 Conceptual framework

The structure of the predicting system is shown in Fig. 1. As can be seen, it comprises three distinct component subsystems, each of which implements the functionality of: triple extraction, link analysis and fuzzy inference, respectively. These activities are integrated to construct a required predicting model, whose implementation steps are detailed below.

2.2 Connected-triple extraction

2.2.1 Concept of connected-triple

Connected-triple modelling, first introduced to analyse global clustering coefficient (Luce and Perry 1949), is also referred to as a method for measuring network transitivity. For instance, it may be applied to measure the extent to which a

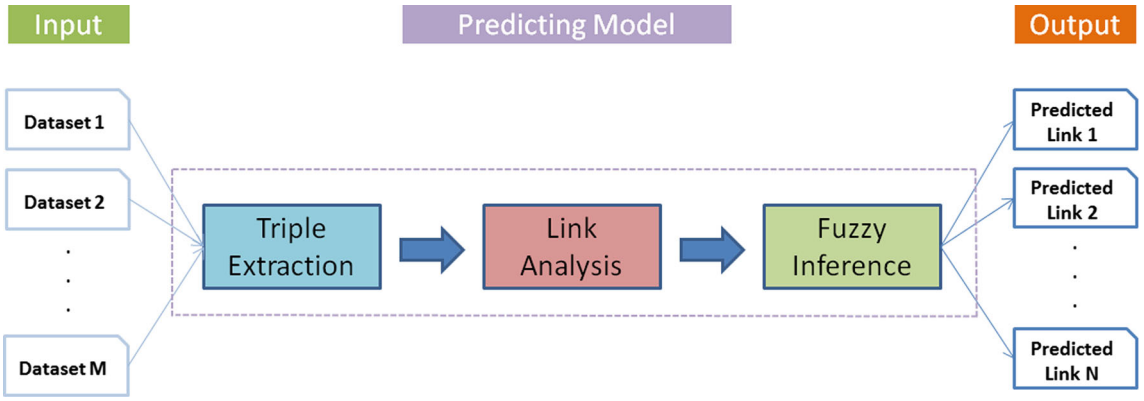
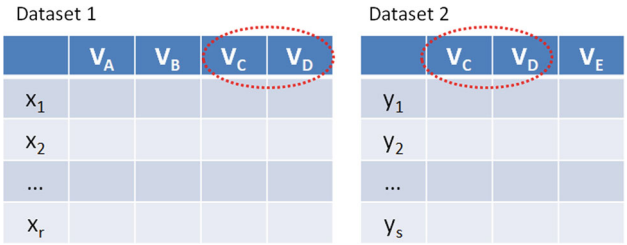


Fig. 1 Predicting framework



	V_A	V_B	V_C	V_D
x_1				
x_2				
...				
x_r				

	V_C	V_D	V_E
y_1			
y_2			
...			
y_s			

Fig. 2 Sample datasets

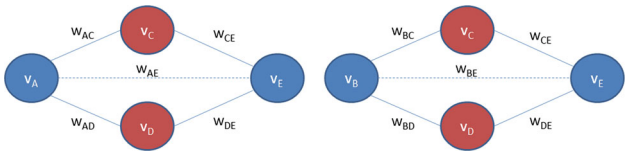


Fig. 3 Connected-triples extracted from sample datasets

friend of someone's friend is also the friend of that person. Formally, a connected-triple, Triple $= \{V_{\text{Triple}}, W_{\text{Triple}}\}$, is a subgraph of $G(V, W)$, where V represents the set of vertices in the graph and W represents the set of edges connecting related pairs of vertices, containing three vertices $V_{\text{Triple}} = \{v_i, v_j, v_k\} \subset V$ and two edges $W_{\text{Triple}} = \{w_{ij}, w_{jk}\} \subset W$, with $w_{ik} \notin W$. The vertex v_j connecting the other two vertices is called the centre of the triple, and v_i or v_k is called an end of the triple (there being two ends per triple, of course).

2.2.2 Extracting connected-triples from datasets

Extracting connected-triples from (the same or different) original datasets plays a fundamental role in the present work. An example of two distinct datasets is shown in Fig. 2, where the variables v_C and v_D co-occur in both datasets (encircled in dot line), while the variables v_A and v_B only appear in Dataset 1, and v_E only appear in Dataset 2. Importantly, an obvious but crucial point is that although there exist variables co-occurring in more than one dataset, these datasets cannot be easily merged into one since the instances in the datasets can be totally distinct and so can the numbers of instances in the datasets. For example, the instances x_1, x_2,..., x_r in Dataset 1 and the instances y_1, y_2,..., x_s in Dataset 2 are completely different from each other, although they share the two aforementioned common variables. Also, Dataset 1 has r instances but Dataset 2 contains s instances, while $r \neq s$.

An example of extracting connected-triples from original datasets is shown in Fig. 3, with each vertex representing a variable in the sample datasets given in Fig. 2. For instance, v_A in Fig. 3 denotes the (same) variable v_A in Dataset 1 of Fig. 2. A link (represented in a solid line) between two distinct variables denotes that these variables are co-occurring in at least one of the sample datasets, and therefore, that they are to a certain extent related to each other. In Fig. 3, four triples, Triple_i, $i = 1, 2, 3, 4$, are formed from Datasets 1 and 2 in Fig. 2, where $V_{\text{Triple\_1}} = \{v_A, v_C, v_E\}$, $V_{\text{Triple\_2}} = \{v_A, v_D, v_E\}$, $V_{\text{Triple\_3}} = \{v_B, v_C, v_E\}$ and $V_{\text{Triple\_4}} = \{v_B, v_D, v_E\}$. The centre of these four connected-triples is v_C and v_D, respectively. The dash line between v_A and v_E and that between v_B and v_E represent the potential links between pairs of the variables v_A and v_E and those of v_B and v_E, respectively, which do not exist in the given datasets.

2.2.3 Transitivity property of connected-triple

An interesting but important characteristic of connected-triple is its transitivity property. According to this property, two independent connected-triples can form a third connected-triple. For instance, as shown in Fig. 4, from $\text{Triple}_1 = \{\{v_1, v_2, v_3\}, \{w_{12}, w_{23}\}\}$, a new link w_{13} connecting v_1 and v_3 may be generated. Likewise, from $\text{Triple}_2 = \{\{v_3, v_4, v_5\}, \{w_{34}, w_{45}\}\}$, another new link w_{35} connecting v_3 and v_5 may also be obtained. Based on the variables v_1, v_3, v_5, and the links w_{13}, w_{35}, an extended connected-

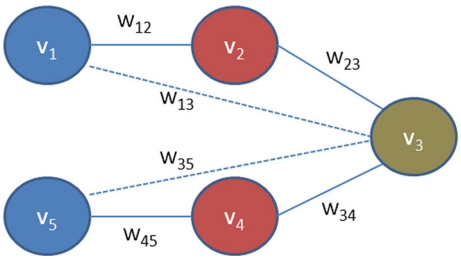


Fig. 4 Transitivity of connected-triple

triple $\text{Triple}_3 = \{\{v_1, v_3, v_5\}, \{w_{13}, w_{35}\}\}$ (depicted with dash lines) can be artificially produced.

2.3 Link analysis

Having identified a new connected-triple from the source datasets, the task of determining correlation between a pair of variables that belong to two different datasets becomes to predict whether there exists a (hidden) link between the two end vertices. If so, a further question is what may be the strength on such a link. To address these issues, prerequisites including the properties of any known links between pairs of vertices in the triple need to be obtained in advance.

In practice, the link property is generally described by its weight, which may correspond to a wide variety of aspects depending on the underlying application problem. For each connection between a given pair of distinct variables, different mechanisms may therefore be devised for estimating the strength of that connection. For instance, in a route graph or map, the weight of a link may indicate the route distance between two venues. In a graph of co-authorship, the weight of a link may denote the number of papers two authors collaborated to publish. In a graph of webpage linkages, the weight on a link may represent the popularity of people stepping from one to another. In the current study, a link between two vertices signifies a certain relationship between those variables in the datasets. Thus, the weight of a link is utilised to capture and reflect the closeness or correlation of the corresponding variables.

2.3.1 Categorical data

For a pair of variables in a dataset filled with discrete or nominal values, their relationship can be described by the co-occurrence frequency of the variables taking on different value-pairs. For such data, two indices to measure link strengths can be adopted, namely normalised mutual information (NMI) and Frequency of Most Popular Term-Pair (FMTP). These strengths are detailed below, which can themselves be combined to form fused link properties.

(1) **Normalised Mutual Information (NMI)** Generally speaking, mutual information is a symmetric measure to quantify the statistical information shared between two distributions (Cover and Thomas 2012). The use of this measure in the present research provides a sound indication of the shared information between a given pair of variables. In particular, for two discrete random variables v_A and v_B, the mutual information between them can be denoted as $MI(v_A, v_B)$ and computed by

$$\text{MI}(v_A, v_B) = \sum_{v_b \in D_B} \sum_{v_a \in D_A} p(v_a, v_b) \log \left(\frac{p(v_a, v_b)}{p(v_a) p(v_b)} \right) \tag{1}$$

where $p(v_a, v_b)$ is the joint probability distribution function of v_A and v_B and $p(v_a)$ and $p(v_b)$ are the marginal probability distribution functions of v_A and v_B, with v_A and v_B defined over the domains D_A and D_B, respectively. Note that there is no upper bound for $MI(v_A, v_B)$. Thus, for better facilitating interpretation and comparison, a normalised version of $MI(v_A, v_B)$ that ranges from 0 to 1 is desirable while describing the relationship strength between v_A and v_B.
Let $H(v_A)$ denote the entropy of v_A (Liang and Shi 2004), which is defined by

$$H(v_A) = -\sum_{v_a \in D_A} p(v_a) \log p(v_a) \tag{2}$$

From this, the normalised mutual information between v_A and v_B (Strehl and Ghosh 2002), denoted by NMI (V_A, V_B), can be computed such that

$$\text{NMI}(v_A, v_B) = \frac{MI(v_A, v_B)}{\sqrt{H(v_A) H(v_B)}} \tag{3}$$

The time complexity of computing NMI is $O(mnd)$, where d denotes the number of instances in the dataset, and m and n represent the cardinalities of variable domains of v_A and v_B, respectively. Typically, m and n are fixed to a small or medium number in advance. From psychological viewpoint, to ensure model interpretability, the cardinalities are normally set to a maximum value of 9. Therefore, this measurement has the linear time complexity proportional to the size of the dataset, namely $O(d)$.

(2) **Frequency of Most Popular Term-Pair (FMPT)** NMI may be a simple measurement computationally. However, only taking it into consideration when modelling the link strengths between distinct variables may not be sufficiently effective. In particular, the frequency of occurrence of different terms with regard to a certain variable within a given dataset can be rather different.

This is because datasets may be rather skewed; certain terms may have a very high occurrence frequency but one or more of the others may have a very low frequency. This is rather common a phenomenon in real-world problems. For example, more than 90% of the primary school pupils are guarded by their parents and they are much less likely to be guarded by other relatives. The statistics of blood type distribution in the UK also show that 44% of the population have blood type O, and only 10% have blood type B (Reid et al. 2012).

When considering any link relationship between two variables v_A and v_B of such skewed datasets, suppose that V_A^1 and V_B^1 are the most popular terms taken by the variables v_A and v_B, respectively. Then, even if most of the instances have the term V_A^1 for v_A and V_B^1 for v_B simultaneously, the NMI score of the link between v_A and v_B may still be low. This is because the NMI score is significantly affected by the number of other term-pairs and their proportion. In this case, judging the link strength between these two distinct variables by only calculating the NMI score may seriously distort the result, misinterpreting the closeness of the relationship between the two. This calls for the development of the so-called frequency of the most popular term-pair measure (FMPT).

Without losing generality, assume that a given dataset includes a total of d instances and that v_A and v_B are two discrete variables describing the instances in the dataset, each containing m and n terms, respectively. Let V_A^i $(1 \leq i \leq m)$ and V_B^j $(1 \leq j \leq n)$ be the terms possibly taken by v_A and v_B, and $S_{V_A^i}$ and $S_{V_B^j}$ $(1 \leq j \leq n)$ be the set of instances which has the term V_A^i for v_A and V_B^i for v_B. The FMPT score or weight on the link between the variable v_A and v_B is defined by

$$\mathrm{FMPT}(v_A, v_B) = \frac{\max\limits_{1\leq i\leq m, 1\leq j\leq n} \mathrm{d}_{(S_{V_A^i} \bigcap S_{V_B^j})}}{\mathrm{d}} \tag{4}$$

where $d_{(S_{V_A^i} \bigcap S_{V_B^j})}$ denotes the number of instances which have the term V_A^i for the variable v_A and V_B^j for v_B simultaneously.

Note that the FMPT score is also ranged from [0,1]. The time complexity of computing FMPT is also $O(mnd)$, where m, n, d are of the same meanings as previously defined.

(3) **Fusion of Link Properties** As indicated above, both NMI and FMPT take values from the same range [0,1]. It is therefore convenient to aggregate the results if both are applied. The fusion of these two measurements is useful because they capture different underlying relationship properties of the datasets in general and the variables' terms in particular. For a certain link between two distinct discrete variables v_A and v_B, given the NMI and FMPT scores, the combined weight of the link $SYN(v_A, v_B)$ can be calculated in a straightforward manner such that

$$\mathrm{SYN}(v_A, v_B) = max(\mathrm{NMI}(v_A, v_B), \mathrm{FMPT}(v_A, v_B)) \tag{5}$$

Obviously, the combined link weight has the same real value range as either of the component weights, i.e. between 0 and 1. The complexity of this fusion step is extremely simple, being $O(2)$. This may be linearly generalised if there are more than 2 such base link strengths. The benefit of adopting the maximum operator is that it takes into consideration the most salient feature of the data while being simple in computation.

Note that the strength fusion does not have to be implemented as above, but can be done in various alternative ways, e.g. by finding the arithmetic average of the component strengths, if preferred. However, this does not affect the approach taken, rather than adding a small amount of extra computational expense and so is regarded as being beyond the scope of the current investigation.

2.3.2 Continuous numeric data

For a pair of variables with continuous data as their entities, the aforementioned measurements may not work. Instead, statistical means to measure the bivariate correlation may be a fitted alternative. Specifically, Pearson correlation coefficient (PCC) (Stigler 1989), a measure of the linear correlation between two continuous variables, is adopted here. A simple but important factor needs to be noted is that for traditional use of PCC, it has a value range between $[-1, 1]$. Considering the main concern here is whether two variables have strong correlation and if so, how strong such a relationship may hold, whether the two variables have a negative or positive correlation is beyond the current scope. Hence, only the absolute value of Pearson correlation coefficient, APCC, is herein employed to measure the link strength between two continuous variables. Formally, the APCC between two variables v_A and v_B can be written as follows:

$$\mathrm{APCC}(v_A, v_B) = \frac{\left| \sum\limits_{g=1}^{d} (V_a^g - \overline{V_A})(V_b^g - \overline{V_B}) \right|}{\mathrm{d}\sigma_{V_A}\sigma_{V_B}} \tag{6}$$

where V_a^g and V_b^g represent the value of v_A and that of v_B for the g-th instance in the dataset, respectively; $\overline{V_A}$ and $\overline{V_B}$ stand for the average value of all the instances with regard to v_A and that to v_B; and σ_{V_A} and σ_{V_B} denote the standard

deviation of v_A and that of v_B within the discussed dataset. The time complexity of computing APCC for any variable pair is $O(d)$, where d denotes the number of instances in the dataset.

2.4 Fuzzy inference model

Having determined the weights over given links within a connected-triple model, the predicting system reaches its final step: logic deduction. A fuzzy inference model is employed to implement this task, providing a flexible means to perform human-interpretable reasoning by the use of linguistic terms rather than numeric values (although the linguistic terms still have their underlying numerical interpretations). For the problem of link prediction, linguistic labels such as "strong", "medium" and "weak" are natural words that are commonly used to describe link strengths. The present work follows this practical observation and attempts to learn the (hidden) links between the network nodes that may be represented in the conventional production rule format:

$$\begin{aligned}&\textbf{IF } \text{link}_1 \textbf{ IS } (\text{strong}\backslash\text{medium}\backslash\text{weak})\\&\textbf{AND } \text{link}_2 \textbf{ IS } (\text{strong}\backslash\text{medium}\backslash\text{weak})\\&\textbf{THEN } \text{link}_3 \textbf{ IS } (\text{strong}\backslash\text{medium}\backslash\textit{weak})\end{aligned} \tag{7}$$

where link_1 and link_2 represent the two known links in a certain triple, each of which connects the triple centre to one of the two ends, and link_3 represents the link to be established with a (predicted) link strength score. Such a fuzzy system involves two key procedures as detailed below.

2.4.1 Link weight fuzzification

To enable the capture and representation of imprecisely described link weights, and to support the derivation of the required fuzzy inference model through data-driven learning, fuzzification of the link strengths for each identified connected-triple is necessary. Without losing generality, to ensure interpretability of the resulting model, a set of membership functions used to depict link strengths is presumed to have been prescribed by domain experts. However, for applications where there is a sufficient amount of historical data, a clustering method may be employed to derive the required set of (potentially more objective) linguistic terms. In this work, especially for the experimental evaluation to be presented in the next section, the linguistic terms used are predefined by the domain experts (with prescribed asymmetrical membership functions used to partition the underlying problem domains), without any optimisation and are shown in Fig. 5.

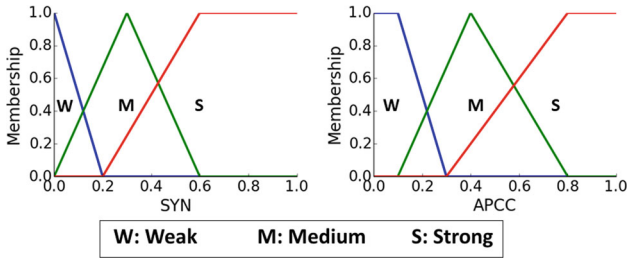


Fig. 5 Fuzzy membership values of link weight with respect to different measures

2.4.2 Fuzzy inference

In the process of performing fuzzy inference for link prediction, as with other applications of fuzzy systems, *t-norm* and *t-conorm* operators are adopted to interpret logic connectives over connected-triples, aggregating fuzzy values (Deschrijver et al. 2004). In general, for each pair of end vertices, there may exist several distinct centres connecting them to form different connected-triples. As such, each connection will lead to an intermediate inference outcome regarding the link strength, indicating the level that that triple may contribute towards the final prediction result. Thus, a *t-conorm* operator is needed to aggregate all the intermediate predicted outcomes together.

Given a connected-triple CT, let $f^L_{\text{link}_1}$ and $f^L_{\text{link}_2}$ be the fuzzy membership values of the link strengths, or link weights on the links link_1 and link_2, where linguistic terms $L \in \mathcal{L}$, with $\mathcal{L}$ representing a collection of all fuzzy sets used to express the linguistic labels (namely, the terms "strong", "medium" and "weak" as given in the previous example). The predicted fuzzy value of a single connected-triple can then be described as a membership function:

$$\begin{aligned}F_{P_{\text{CT}}} = \Big[&\nabla(f^{L_1}_{\text{link}_1}, f^{L_1}_{\text{link}_2}), \quad \nabla(f^{L_2}_{\text{link}_1}, f^{L_2}_{\text{link}_2}), \ldots,\\ &\nabla(f^{L_M}_{\text{link}_1}, f^{L_M}_{\text{link}_2})\Big]\end{aligned} \tag{8}$$

where ∇ denotes a certain predefined *t-norm* and M represents the number of the terms possibly used to describe the linguistic link strength.

Suppose that there are N connected-triples formed by a specific pair of end vertices with a common corresponding centre, the predicted fuzzy value for the link strength of P_{link} can be logically interpreted as follows:

$$\begin{aligned}F_{P_{\text{link}}} = [&\Delta(f^{L_1}_{P^1_{\text{CT}}}, f^{L_1}_{P^2_{\text{CT}}}, \cdots, f^{L_1}_{P^N_{\text{CT}}}),\\ &\Delta(f^{L_2}_{P^1_{\text{CT}}}, f^{L_2}_{P^2_{\text{CT}}}, \cdots, f^{L_2}_{P^N_{\text{CT}}}),\\ &\vdots\end{aligned}$$

$$\Delta(f_{P_{\mathrm{CT}}^1}^{L_M}, f_{P_{\mathrm{CT}}^2}^{L_M}, \cdots, f_{P_{\mathrm{CT}}^N}^{L_M})] \tag{9}$$

where Δ represents an extended version of a certain *t-conorm* which can take a finite number of arguments. It aggregates those fuzzy membership values obtained from each connected-triple corresponding to a pair of variables and generates a new fuzzy membership value for the predicted link between those two variables. As the final result, what is returned is a fuzzy value regarding that to what extent a detected link is of a certain strength with respect to each individual predefined link weights (that are provided by the domain experts). If, however, it is desirable to provide a numerical number for the predicted link strength, an additional computational step is to defuzzify the resulting fuzzy membership value.

2.5 Illustrative link strength prediction

Consider two simple datasets regarding student academic performance, as shown in Fig. 6. The two datasets contain 14 and 10 distinct instances, respectively, with the attributes "1st semester grade" and "2nd semester grade" shared by both. This illustrative example is to demonstrate that the proposed approach can predict the correlation between the variable "Family support" in Dataset 1 and the variable "Family size" in Dataset 2, with a measured link strength.

For shorthand, denote the variables "Family support", "1st semester grade", "2nd semester grade" and "Family size" as "fsup", "1sg", "2sg" and "fzise", respectively. Then, from the given datasets, the following two connected-triples can be directly extracted from these datasets, one from each:

$$\text{Triple\_1} = \{\{v_{\text{fsup}}, v_{\text{1sg}}, v_{\text{fsize}}\}, \{w_{\text{fsup}-\text{1sg}}, w_{\text{1sg}-\text{fsize}}\}\}$$
$$\text{Triple\_2} = \{\{v_{\text{fsup}}, v_{\text{2sg}}, v_{\text{fsize}}\}, \{w_{\text{fsup}-\text{2sg}}, w_{\text{2sg}-\text{fsize}}\}\}$$

From these, according to formulae (1), (2) and (3) it can be computed that:

$$\text{NMI}(V_{\text{fsup}}, V_{\text{1sg}}) = 0.139, NMI(V_{\text{fsup}}, V_{\text{2sg}}) = 0.172$$
$$\text{NMI}(V_{\text{fsize}}, V_{\text{1sg}}) = 0.580, NMI(V_{\text{fsize}}, V_{\text{2sg}}) = 0.474$$

Similarly, through formula (4) it can be computed that:

$$\text{FMPT}(V_{\text{fsup}}, V_{\text{1sg}}) = 0.429, \text{FMPT}(V_{\text{fsup}}, V_{\text{2sg}}) = 0.357$$
$$\text{FMPT}(V_{\text{fsize}}, V_{\text{1sg}}) = 0.300, \text{FMPT}(V_{\text{fsize}}, V_{\text{2sg}}) = 0.300$$

Thus, the weights on these links can be computed by formula (5), such that

$$\text{SYN}(V_{\text{fsup}}, V_{\text{1sg}}) = \max(0.139, 0.429) = 0.429$$
$$\text{SYN}(V_{\text{fsup}}, V_{\text{2sg}}) = \max(0.172, 0.357) = 0.357$$
$$\text{SYN}(V_{\text{fsize}}, V_{\text{1sg}}) = \max(0.580, 0.300) = 0.580$$
$$\text{SYN}(V_{\text{fsize}}, V_{\text{2sg}}) = \max(0.474, 0.300) = 0.474$$

Having acquired the weights for the existing links (within individual datasets), the next step is to conduct fuzzy inference. Suppose that the fuzzy membership functions of a synthesised link strength are provided by the domain experts in linguistic terms as specified in Fig. 5. In this simple illustration, assume that the Max-Min aggregation method is taken to compute the SYN weights. Then, for Triple_1, according to formula (8), its weight $F_{P_{\mathrm{CT}}}$ can be calculated such that

$$[\min(f^W(0.429), f^W(0.58)), \min(f^M(0.429), f^M(0.58)),$$
$$\min(f^S(0.429), f^S(0.58))] = [0, 0.067, 0.5725]$$

where f^W, $f^M and f^S$ denote the fuzzification results of the link weights with respect to the linguistic terms "weak", "medium" and "strong", respectively. What this fuzzy result stands for is that the detected link is not "weak" (as it is of a zero membership value with regard to this strength label), a tiny membership for the fuzzy concept "medium" and a significant membership value for the given linguistic term "strong". Following the same calculating procedure, for Triple_2, its $F_{P_{\mathrm{CT}}}$ score is [0, 0.42, 0.3925]. Hence, with respect to formula (9), the predicted fuzzy value representing the strength of the link between variables "Family support" and "Family size" is:

$$[\max(0, 0), \max(0.067, 0.42), \max(0.5725, 0.3925)]$$
$$= [0, 0.42, 0.3925]$$

Finally, if a numerical strength score between the two variables is desirable (as opposite to a fuzzy value), then by employing the centre of gravity (COG) method for defuzzification, the predicted link score of 0.5626 can be obtained in a straightforward manner. Note that the above illustrative example is carried out on categorical datasets. However, this method can also be applied to numeric datasets, although the strategy to measure link strengths needs to be adjusted accordingly.

3 Empirical evaluation

3.1 Datasets

The experimental evaluation is conducted on both real-world data from UCI benchmark datasets (Bache and Lichman 2013) and on a collection of synthetic datasets. Since there is hardly any corpora of datasets designed particularly for

Sample data 1

No	Family support	1st semester grade	2nd semester grade
1	yes	B	B
2	yes	A	B
3	yes	B	A
4	no	C	C
5	yes	B	A
6	no	B	C
7	yes	A	B
8	yes	B	B
9	no	C	B
10	no	B	B
11	yes	B	B
12	yes	C	C
13	yes	B	A
14	no	B	B

Sample data 2

No	Family size	1st semester grade	2nd semester grade
1	large	B	C
2	medium	B	B
3	small	A	B
4	small	A	A
5	medium	B	B
6	large	C	C
7	large	C	B
8	small	A	A
9	small	B	B
10	medium	B	B

Fig. 6 Two simple datasets used for illustration

the current study, the datasets from UCI benchmark are split into several subsets with overlapped variables according to human knowledge. To test the performance of the proposed approach on larger sized groups of datasets involving more variables, four different corpora of synthetic datasets are also generated to conduct the experiment. Table 1 shows a summary of the characteristics of all datasets employed.

Note that as with any real-world application, the ground truth of the link strengths between variables is not a natural existence in these datasets. Thus, in the following experiments, any "ground truth" is artificially computed by the testing data using the corresponding method as outlined in Sect. 2.3. That is, without losing fairness, the predicted results are compared against those directly generated by the same underlying link strength measurement from the testing data. This may sound unintuitive, but it allows for fair comparisons to be carried out.

3.2 Methods for comparison

Predicting link strengths among variables observed in different data sources is a brand new topic. As such, it is impractical to directly compare this work with any existing work with respect to this novel problem. Instead, a set of existing link prediction methods based on graph topology are implemented for comparison. In each of the compared methods, each variable is regarded as a node in the graph, and the similarity score between two variables is interpreted as the weight of the assumed link for the corresponding pair of nodes.

3.2.1 Neighbour-based metrics

– **Weighted Jaccard Coefficient (WJC)** This metric assesses and normalises the weights of the common neighbours of a given pair of variables (Dimitriadou et al. 2004). It assigns higher values for pairs of variables which share a higher sum of weights over common neighbours relative to the total weights of all the neighbours they have. For two distinct variables v_A and v_B, this measure is defined by

$$\mathrm{WJC}(v_A, v_B) = \frac{\sum_{v_C \in \Gamma(v_A) \bigcap \Gamma(v_B)} w(v_A, v_C) + w(v_B, v_C)}{\sum_{v_D \in \Gamma(v_A) \bigcup \Gamma(v_B)} w(v_A, v_D) + w(v_B, v_D)}$$

where $\Gamma(v_A)$ and $\Gamma(v_B)$ denote the adjacent neighbourhood of v_A and that of v_B, respectively.

– **Weighted Resource Allocation (WRA)** This metric is motivated by the physical process of resource allocation (Zhou et al. 2009). Different from WJC, WRA does not only involve adjacent neighbours, but also exploit the neighbours of the direct neighbours of a pair of variables. Formally, WRA is defined as:

$$\mathrm{WRA}(v_A, v_B) = \sum_{v_C \in \Gamma(v_A) \bigcap \Gamma(v_B)} \frac{w(v_A, v_C) + w(v_C, v_B)}{s(v_C)}$$

where $s(v_C)$ denoted the sum of weights for the variable v_C associated with all of its existing links, such that

$$s(v_C) = \sum_{v_X \in \Gamma(v_C)} w(v_C, v_X)$$

3.2.2 Path-based metrics

– **Local Weighted Path (LWP)** This metric is an extended version (Thi et al. 2014) of the local path (LP) method (Lü et al. 2009); it is also a special case of the well-known Katz algorithm (Katz 1953). Unlike the metrics

Table 1 Summary of datasets

Dataset collection	Type	NDC	ANIED	ANVED	ANVCTD
Bank	C	7	6459	4	2
Mushroom	C	9	912	5	2
Salary	C	6	5028	6	1
Student-Por	C	4	163	8	3
Student-Mat	C	4	101	8	2
Connect-4	C	6	11034	7	2
Wine	N	4	1256	3	1
Twitter	N	7	82038	13	4
Facebook	N	5	104	5	2
Urban	N	13	203	12	3
News	N	12	3048	7	2
Music	N	11	211	11	4
Synthetic-1	C	22	15491	18	4
Synthetic-2	C	30	24823	20	3
Synthetic-3	N	25	21990	20	2
Synthetic-4	N	35	19926	18	4

C categorical data, *N* numeric data, *NDC* number of datasets in collection, *ANIED* average number of instances in each dataset, *ANVED* average number of variables in each dataset, *ANVCTD* average number of variables coexisting in two datasets

that only use the information of the nearest neighbours (be they adjacent or otherwise), LWP makes use of further information from local paths with a length value of 2 and 3. Let A denote the weighted adjacent matrix of all variables under discussion, and A^2 and A^3 represent the weighted adjacent matrices based on A with a length of 2 and 3, respectively, LWP is then defined as follows:

$$\text{LWP}= A^2 + \alpha A^3$$

where α is a small number close to 0, which is being used to penalise the weight of the paths with greater length. In the experiment, α is set to 0.01 (as with the default value typically used when running this metric).

– **Relation Strength Similarity (RSS)** This metric was originally introduced as an asymmetric measure for weighted social networks (Chen et al. 2012). It may also be adopted as a symmetric measure for the problem considered herein. For the present study, suppose that there are T simple paths shorter than a path length of e from the variable v_A to v_B, and a path with length of u ($u \leq e$) from v_A and v_B is formed with Z variables v_1, v_2,…, v_{Z-1}, v_Z, where v_1 represents v_A and v_Z represents v_B. Then, the RSS metric from v_A to v_B is defined by

$$\text{RSS}(v_A, v_B) = \sum_{t=1}^{T} R_u^*(v_A, v_B)$$

with

$$R_u^*(v_A, v_B) = \begin{cases} \prod_{z=1}^{Z-1} R(v_z, v_{z+1}) & Z \leq e+1 \\ 0 & \text{otherwise} \end{cases}$$

$R(v_z, v_{z+1})$ denotes the link strength of the two adjacent variables v_z and v_{z+1} within a particular path connecting v_A and v_B. Note that for the present experiment, in order to guarantee that each pair of variables have at least one path connecting them, e is set to $k-1$, where k is the number of datasets in the corpus.

3.2.3 Random walk-based metric

– **SimRank (SR)** The well-known SimRank algorithm (Jeh and Widom 2002) was proposed on the basis of the intuition that two nodes within a graph are similar if they are connected to similar nodes in the graph. This can obviously be adapted for use in the present study. For a pair of variables v_A and v_B, their SR score is computed by

$$\text{SR}(v_A, v_B) = \frac{\gamma \sum_{v_C \in \Gamma(v_A)} \sum_{v_D \in \Gamma(v_B)} SR(v_C, v_D)}{|\Gamma(v_A)||\Gamma(v_B)|}$$

where the constant $\gamma \in [0, 1]$ is a decay factor, representing the confidence level of accepting two non-identical variables as similar. In this work, γ is set to 0.8, as being widely used

in various applications. Additionally, the number of iterations for running the SR algorithm is set to 20 in the present experimental evaluation.

3.3 Experimental setup

In all experiments carried out, for each corpus of datasets, an n-fold cross-validation (Bengio and Grandvalet 2005) is performed, where n is the number of datasets in each corpus. The following reported results are based on an average of running 10 times n-fold cross-validation.

Note that not all the approaches implemented for comparison may necessarily generate predicted results ranging between [0, 1], and normalising these results can only provide relative values for all the predicted links, thereby possibly misleading the interpretation of the computed link strength. A precision measure which calculates the percentage of correct predictions according to the portion of founded links is therefore employed to articulate the predicting accuracy. In particular, for all unobserved links defined by the training datasets, their predicted link strengths are computed by each of the predicting algorithms and then ranked in descending order. Simultaneously, their link strengths are calculated through the testing datasets and ranked in descending order as well. The predicating accuracy is determined by comparing the number of the correct predictions against the scenario where a specific portion of unobserved links is assumed.

When conducting the experiments, for simplicity and clarity, $t\text{-}norm$ and $t\text{-}conorm$ are initially implemented with minimum and maximum operators, respectively. To reflect the flexibility of the proposed approach, and also to strengthen comparative studies, another type of operator combination, namely algebraic product and bounded sum, are also applied to form the bounded sum algebraic product (BSAP) interpretation. Additionally, the centre of gravity (COG) method is employed to perform in the defuzzification step.

3.4 Experimental results

The experimental results are measured by predicting accuracy, that is, the ratio of the number of correctly predicted results that are disclosed by each compared method, over the percentage of retrieved variable pairs. In experimental running, all potential variable pairs are examined and ranked in descending order, and the top-K percent of the disclosed variable pairs is selected to compare against the "ground truth" (as indicated in Sect. 3.1) with the corresponding ratio. The predicting results revealed in this paper are simply based on the top-ranked variable pairs within 50% of them all. This is because the predicting accuracy generally retains an increasing trend in response to the increasing ratio of predicted links. When the number of predicted links reaches its maximum, meaning that all potential variable pairs are taken into account, the predicting accuracy will be 1. In reality, it is the highly ranked variable pairs that are more attractive to both human analysts and the general public.

3.4.1 Experimental results for real data

Figure 7 and Fig. 8 show the experimental results for the corpora of real-world datasets. These results jointly demonstrate that the proposed approach is generally very competitive under different circumstances. In particular, the proposed method consistently outperforms the neighbour-based metrics (WJC, WRA) and the random walk-based metric (SimRank) across most corpora of datasets. Note that LWP and RSS metric can perform well for specific corpora of datasets. This has much to do with the distribution of the variables in each dataset within such a corpus. For instance, in the corpus of salary datasets, most of the variables have an explicit relation with the variable "salary", which is natural and makes it easy for the LWP metric to handle. This is also the case for the corpora of Student-Por, Student-Mat and Music datasets. In the corpus of Twitter datasets, almost each dataset is simply connected with only one another through a few number of overlapped variables, which is suitable for RSS, but may not fit for others such as neighbour-based metrics. However, the proposed method is still competitive according to the predicting accuracy, regardless of the variable distribution for each dataset corpus. This shows the robustness and adaptability of the present approach.

Another interesting finding is that for variable pairs with the "strongest" link strength, say, top 5% or top 10%, the proposed approach performs best among all compared methods to identify them. Although the predicting accuracy is not sufficiently high to meet human expectation, it is worth recognising that the task of identifying "strong" links is much more difficult than just finding whether there is a general link (Boongoen et al. 2010). Such a detection is also of practical significance since in real-world applications, it is the identification of any variable pair that is associated with the most "strong" link that is generally more attractive to the users.

Note that for the two distinct implementations of the proposed approach, using either Max-Min or BSAP interpretation, it is difficult to judge which one performs better. This may reflect the robustness of the underlying approach, but no theoretical proof for this hypothesis is done, which remains as active further research.

3.4.2 Experimental results for synthetic data

Another set of experimentations has been conducted on different corpora of synthetic datasets. The experimental results are shown in Fig. 9. It can be seen that the predicting accu-

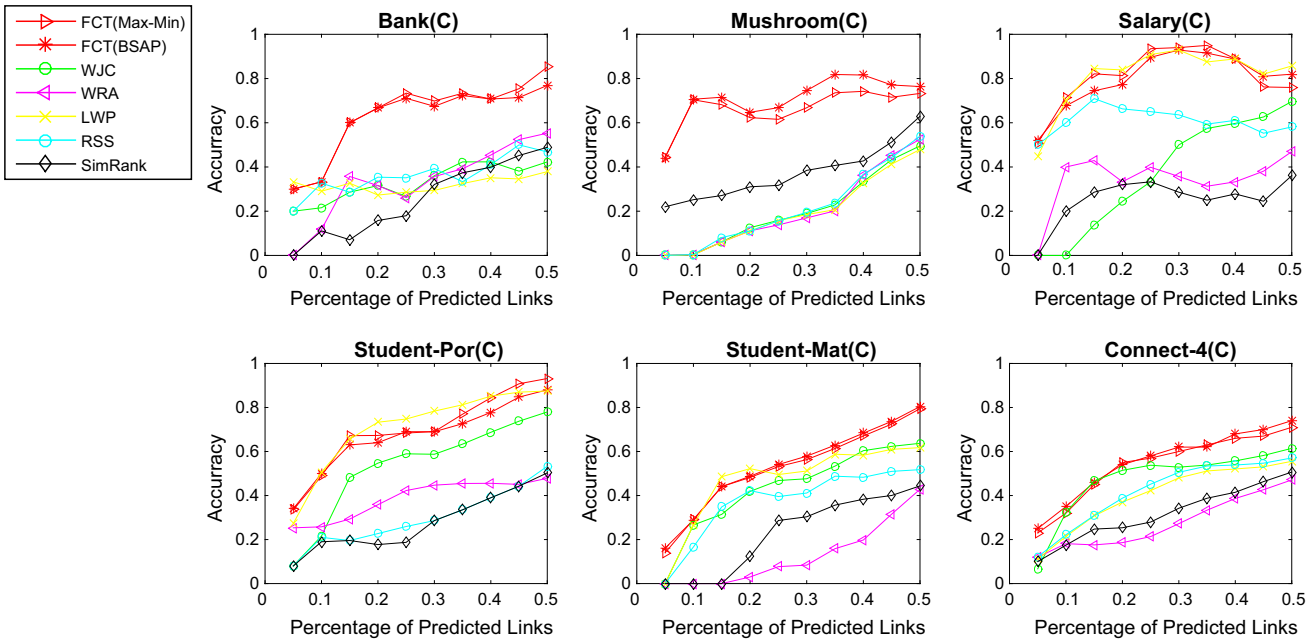


Fig. 7 Prediction accuracy for real-world categorical data

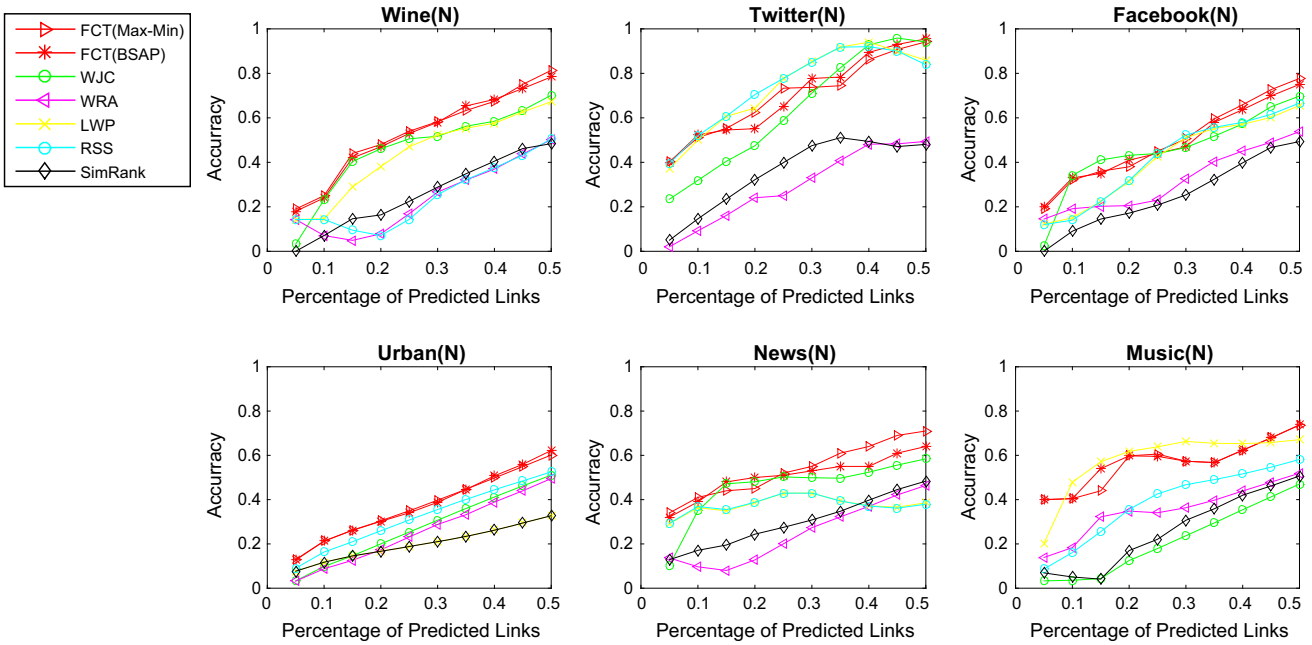


Fig. 8 Prediction accuracy for real-world continuous numeric data

racy for several of the compared methods declines to an extent with respect to the results shown earlier. Due to the increasing number of datasets in the corpus, certain incorporated datasets may not necessarily have common variables amongst them. This situation makes accurate prediction far more difficult.

The neighbour-based metrics suffer significantly from this condition. Both WJC and WRA lead to unsatisfactory results. This may be expected by the fact that neither of them is able to make integrated consideration of both nearest neighbour and non-nearest neighbour variables across different datasets. SR performs slightly better than neighbour-based metrics on lager corpus of datasets, as it does not take just common neighbour variables into account. The propagation of similarity scores amongst variable pairs in the entire dataset corpus could have a positive effect on predicting accuracy for lager

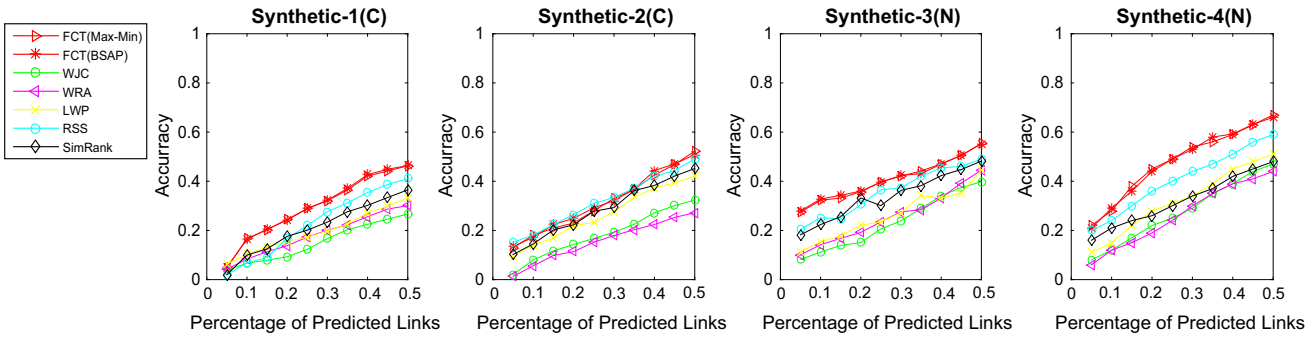


Fig. 9 Prediction accuracy for artificial data

Table 2 Analysis of time complexity

Method	Time complexity
FCT	$O(k^2p^2q^2lf)$
WJC	$O(k^2p^2q^2)$
WRA	$O(k^2p^2q^2)$
LWP	$O(k^3p^3)$
RSS	$O(k^2p^2q^{e+1})$
SR	$O(k^2p^2q^2r)$

l Number of linguistic terms to describe link strength, f time for defuzzification, e path length for RSS, r number of iterations for SR

sized dataset corpus. Interestingly, the performance of RSS has not been adversely much affected by the increased size of dataset collections, since it guarantees to find routes connecting a variable pair. Conversely, the performance of LWP drops dramatically, because it only takes paths of a length of 2 and 3 into consideration. Nevertheless, compared against all these, the proposed approach still performs better in most of the cases, illustrating once again the efficacy of utilising the transitivity property over the structure of connected-triples.

3.5 Complexity analysis

In addition to evaluating these methods in terms of predicting accuracy, it is important to investigate the computational complexity that would determine their actual efficiency for real-world applications. Suppose that a particular corpus incorporates k datasets, that each dataset contains P variables on average, and that every two datasets share on average q identical variables. Table 2 shows the time complexity to find the correlation between all the potential variable pairs for each of the compared algorithms.

WJC and WRA are the most efficient metrics among the compared methods, SR is slightly more expensive than WJC and WRA, with required r iterations of refinement. Generally, the time complexity of path-based metrics is higher than that of neighbour-based methods. LWP has the cubical time complexity since it involves matrix multiplication. The time complexity for RSS is significantly affected by the length of paths and can be extraordinarily high in extreme cases.

Although the proposed approach is not the most satisfactory in terms of time complexity due to the time expenses incurred by performing fuzzy inference, it is acceptable, especially when compared with the path-based metrics dealing with large corpus involving many datasets. Taking both predicting accuracy and time complexity into joint consideration, the proposed approach is considerably competitive upon most occasions.

4 Conclusion and future work

This paper has presented a novel data-driven approach to predict the connections between variables that are hidden in different datasets. Techniques for measuring correlation between domain variables of a certain corresponding type have been proposed. Assisted with the concept of fuzzy connected-triple, the relationships between distinct variables and their transitivity can be naturally captured, represented and reasoned through the link notation. The use of fuzzy inference supports the link prediction process to be more consistent with human reasoning, with the predicted results being readily interpretable. Experimental results on different corpora of datasets have shown that the proposed approach generates more accurate predicted outcomes, while involving relatively simple computation.

While promising, the proposed work is open for further investigation. Within the current implementation, for the step of fuzzy inference, each connected-triple has been treated equally. A better approach might be to aggregate these connected-triples according to the importance of the individual centres of the triples, boosting the reliability of the detected links (Boongoen et al. 2011). Alternative aggregating methods (e.g. arithmetic average and ordered weighted averaging as employed in (Su et al. 2017)) could be utilised

for this. Also, the propagation and aggregation operations developed in social trust networks (Verbiest et al. 2012; Victor et al. 2011a, b) may be adapted for such use to enhance the reliability of the predicting system. Furthermore, the current study conducts exhaustive search for all potential variable pairs; an aided heuristic metric for disclosing variable pairs with "strong" correlation may help to reduce the time complexity. For link strength measurement, metrics other than those presently employed may be considered to further improve the modelling performance. Moreover, in this work, only datasets involving pure categorical values and those involving continuous numeric values are tested. Developing other types of link analysis strategy to handle mixed-type datasets is clearly desirable.

Acknowledgements The first author is grateful to Aberystwyth University for providing a PhD scholarship in support of this research. The authors would also like to thank the reviewers of the initial version (Li et al. 2017) of this paper (which was presented at the 17th Annual Workshop on Computational Intelligence) for their strong support in the work such that it received one of the two best paper awards at the Workshop; their constructive comments on the earlier version have helped improve this work significantly. This study was mainly self-funded; other than the first author receiving a Ph.D. scholarship from Aberystwyth University, no external funding was received in support of this research.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants performed by any of the authors.

References

Aiello LM, Barrat A, Schifanella R, Cattuto C, Markines B, Menczer F (2012) Friendship prediction and homophily in social media. ACM Trans Web (TWEB) 6(2):9

Akcora CG, Carminati B, Ferrari E (2011) Network and profile based measures for user similarities on social networks. In: Information Reuse and Integration (IRI), 2011 IEEE International Conference on, IEEE, pp 292–298

Almansoori W, Gao S, Jarada TN, Elsheikh AM, Murshed AN, Jida J, Alhajj R, Rokne J (2012) Link prediction and classification in social networks and its application in healthcare and systems biology. Netw Model Anal Health Inf Bioinf 1(1–2):27–36

Bache K, Lichman M (2013) Uci machine learning repository

Bae SH, Nikolaev A, Seo JY, Castner J (2015) Health care provider social network analysis: a systematic review. Nurs Outlook 63(5):566–584

Barabâsi AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. Physica A: Stat Mech Appl 311(3–4):590–614

Bengio Y, Grandvalet Y (2005) Bias in estimating the variance of k-fold cross-validation. In: Statistical modeling and analysis for complex data problems, Springer, pp 75–95

Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: structure and dynamics. Phys Rep 424(4):175–308

Boongoen T, Shen Q, Price C (2010) Disclosing false identity through hybrid link analysis. Artif Intell Law 18(1):77–102

Boongoen T, Shang C, Iam-On N, Shen Q (2011) Fuzzy orders-of-magnitude based link analysis for qualitative alias detection. IEEE Trans Syst, Man, Cybern, Part B: Cybern 41:1705–1714

Bringmann B, Berlingerio M, Bonchi F, Gionis A (2010) Learning and predicting the evolution of social networks. IEEE Intell Syst 25(4):26–35

Chen HH, Gou L, Zhang XL, Giles CL (2012) Discovering missing links in networks using vertex similarity measures. In: Proceedings of the 27th annual ACM symposium on applied computing, ACM, pp 138–143

Colladon AF, Remondi E (2017) Using social network analysis to prevent money laundering. Expert Syst Appl 67:49–58

Cover TM, Thomas JA (2012) Elements of information theory. Wiley, London

Deschrijver G, Cornelis C, Kerre EE (2004) On the representation of intuitionistic fuzzy t-norms and t-conorms. IEEE Trans Fuzzy Syst 12(1):45–61

Dimitriadou E, Barth M, Windischberger C, Hornik K, Moser E (2004) A quantitative comparison of functional mri cluster analysis. Artif Intell Med 31(1):57–71

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, Von Mering C et al (2012) String v9. 1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41(D1):D808–D815

Huang Z, Lin DK (2009) The time-series link prediction problem with applications in communication surveillance. INFORMS J Comput 21(2):286–303

Jeh G, Widom J (2002) Simrank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 538–543

Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43

Kim M, Leskovec J (2011) The network completion problem: Inferring missing nodes and edges in networks. In: Proceedings of the 2011 SIAM International Conference on Data Mining, SIAM, pp 47–58

Li Z, Shang C, Shen Q (2017) Fuzzy connected-triple for predicting inter-variable correlation. In: Proceedings of the 17th annual workshop on computational intelligence, pp 49–62

Liang J, Shi Z (2004) The information entropy, rough entropy and knowledge granulation in rough set theory. Int J Uncertain, Fuzziness Knowl-Based Syst 12(01):37–46

Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Assoc Inf Sci Technol 58(7):1019–1031

Lü L, Zhou T (2010) Link prediction in weighted networks: the role of weak ties. EPL (Europhys Lett) 89(1):18,001

Lü L, Jin CH, Zhou T (2009) Similarity index based on local paths for link prediction of complex networks. Phys Rev E 80(4):046–122

Luce RD, Perry AD (1949) A method of matrix analysis of group structure. Psychometrika 14(2):95–116

Marchette DJ, Priebe CE (2008) Predicting unobserved links in incompletely observed networks. Comput Stat Data Anal 52(3):1373–1386

Martínez V, Berzal F, Cubero JC (2017) A survey of link prediction in complex networks. ACM Comput Surve (CSUR) 49(4):69

Mori J, Kajikawa Y, Kashima H, Sakata I (2012) Machine learning approach for finding business partners and building reciprocal relationships. Expert Syst Appl 39(12):10,402–10,407

Newman ME, Park J (2003) Why social networks are different from other types of networks. Phys Rev E 68(3):036–122

Pavlov M, Ichise R (2007) Finding experts by link prediction in co-authorship networks. In: Proceedings of the 2nd International Conference on Finding Experts on the Web with Semantics-Volume 290, Citeseer, pp 42–55

Reid ME, Lomas-Francis C, Olsson ML (2012) The blood group antigen factsbook. Academic Press, New York

Shen Q, Boongoen T (2012a) Extending data reliability measure to a filter approach for soft subspace clustering. IEEE Trans Knowl Data Eng 24:649–664

Shen Q, Boongoen T (2012b) Social network inspired approach to intelligent monitoring of intelligence data. Social Network Mining, Analysis and Research Trends: Techniques and Applications, IGI Publishing 24:79–100

Stigler SM (1989) Francis galton's account of the invention of correlation. Stat Sci pp 73–79

Strehl A, Ghosh J (2002) Cluster ensembles–a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3:583–617

Su P, Shang C, Chen T, Shen Q (2017) Exploiting data reliability and fuzzy clustering for journal ranking. IEEE Trans Fuzzy Syst 25:1306–1319

Thi DB, Ichise R, Le B (2014) Link prediction in social networks based on local weighted paths. In: Future Data and Security Engineering, Springer, pp 151–163

Varone F, Ingold K, Jourdain C, Schneider V (2017) Studying policy advocacy through social network analysis. Eur Political Sci 16:322–336

Verbiest N, Cornelis C, Victor P, Herrera-Viedma E (2012) Trust and distrust aggregation enhanced with path length incorporation. Fuzzy Sets Syst 202:61–74

Victor P, Cornelis C, De Cock M (2011a) Trust networks for recommender systems, vol 4. Springer, Berlin

Victor P, Cornelis C, De Cock M, Herrera-Viedma E (2011b) Practical aggregation operators for gradual trust and distrust. Fuzzy Sets Syst 184(1):126–147

Wang P, Xu B, Wu Y, Zhou X (2015) Link prediction in social networks: the state-of-the-art. Sci China Inf Sci 58(1):1–38

Wasserman S, Faust K (1994) Social network analysis: methods and applications, vol 8. Cambridge University Press, Cambridge

Wasserman S, Galaskiewicz J (1994) Advances in social network analysis: Research in the social and behavioral sciences, vol 171. Sage Publications, Beverley Hills

Wu S, Sun J, Tang J (2013) Patent partner recommendation in enterprise social networks. In: Proceedings of the sixth ACM international conference on Web search and data mining, ACM, pp 43–52

Zech ST, Gabbay M (2016) Social network analysis in the study of terrorism and insurgency: From organization to politics. Int Stud Rev 18(2):214–243

Zheng X, Le Y, Chan AP, Hu Y, Li Y (2016) Review of the application of social network analysis (sna) in construction project management research. Int J Project Manag 34(7):1214–1225

Zhou T, Lü L, Zhang YC (2009) Predicting missing links via local information. Eur Phys J B 71(4):623–630

Zhu J, Hong J, Hughes JG (2002) Using markov models for web site link prediction. In: Proceedings of the thirteenth ACM conference on Hypertext and hypermedia, ACM, pp 169–170